ABSTRACT

Data mining applications have enormously altered the strategic decision-making processes of organizations. The application of association rules algorithms is one of the well-known data mining techniques that have been developed to cope with multidimensional databases. However, most of these algorithms focus on multidimensional data models for transactional data. As data warehouses can be presented using a multidimensional model, in this paper we provide another perspective to mine association rules in data warehouses by focusing on a measurement of summarized data. We propose four algorithms — VAvg, HAvg, WMAvg, and ModusFilter — to provide efficient data initialization for mining association rules in data warehouses by concentrating on the measurement of aggregate data. Then we apply those algorithms both on a non-repeatable predicate, which is known as mining normal association rules, using GenNLI, and a repeatable predicate using ComDims and GenHLI, which is known as mining hybrid association rules.

Keywords: association rules; data mining; data preprocessing; data warehouse; multidimensional database design

INTRODUCTION

Association rules on transaction database were first introduced by Agrawal (1993). By using its Apriori algorithm, large itemsets satisfying the minimum support and association rules based on the minimum confidence could be discovered. Since then, a large number of efficient algorithms using the hash-based technique (Park, Chen, & Yu, 1995), transaction reduction (Han & Fu, 1995), the partition technique (Mannila, Toivonen, & Verkamo, 1994), and sample datasets to prune the number of passes on the data (Toivonen, 1996) have been introduced.

Association rules traditionally use transactional data that focus on a single dimension or predicate (Agrawal & Srikant, 1994; Han & Fu, 1995; Mannila, Toivonen, & Verkamo, 1994; Park, Chen, & Yu, 1995; Savasere, Omiecinski, & Navathe, 1995). However, this is not adequate since real life data usually involves more than one dimension or predicate. Subsequently, traditional association rules were developed to
solve the multidimensional model (Guenzel, Albrecht, & Lehner, 1999; Kamber et al., 1997). Kamber et al. (1997) exposed the idea of mining association rules in a multidimensional data model. Their algorithm focuses only on presenting association rules in a multidimensional model, which involves more than one dimension in transactional data. However, this algorithm did not discuss the hierarchies that are also characteristic of a multidimensional model. Later on, a new algorithm (Guenzel, Albrecht, & Lehner, 1999) was proposed to support mining multidimensional databases by hierarchy using an online analytical processing (OLAP) approach.

Apparently, both concepts in Guenzel, Albrecht, and Lehner (1999) and Kamber et al. (1997) miss the most important attribute, which is the measurement of aggregate data in a Data Warehouse (DW). The data in a DW contains only summarised data such as quantity sold, amount sold, and etcetera. No transactions data is stored. In this paper, we focus on providing a framework for mining association rules both on non-repeatable predicates and repeatable predicates in data warehouses by concentrating on the measurement of aggregate data.

Here, we propose four algorithms — $H_{Avg}$, $V_{Avg}$, $ModusFilter$, and $WMAvg$ — to provide efficient data initialisation for mining association rules in data warehouses by concentrating on the measurement of aggregate data, specifically its quantity. Then we apply those algorithms both to non-repeatable predicates using $GenNLI$, and repeatable predicates using $ComDims$, and $GenHLI$, which is known as mining hybrid association rules.

As shown in Figure 1, we provide a framework for mining association rules in a data warehouse. We use quantity data in a fact table to explain our approach. There are three steps to perform in order to derive an initialised data for mining association rules in a data warehouse. First, we select the data warehouse that we want to mine. Secondly, we use user input variables to decide the dimensions that will be used along with the single or interval data to find the interesting patterns. Finally, we use our approach: $H_{Avg}$, $V_{Avg}$, $ModusFilter$, and $WMAvg$ algorithms to produce data initialisation based on the information derived from the two previous steps (see Figure 1). Then we mine the DW using data initialisation from those proposed algorithms to mine non-repeatable association rules and the repeatable predicate, which is known as mining hybrid association rules.

$H_{Avg}$, $V_{Avg}$ and $WMAvg$ algorithms work by selecting the average measurement of aggregate data in a DW with multidimensional structures such as average quantity sold, average price, and so forth. We prune all the rows in the fact table, which have less than the average quantity, since we assume that rows with quantities less than its average will not form any association rules. The main differences between these algorithms are: $H_{Avg}$ finds the average quantity of the defined dimensions horizontally while $V_{Avg}$ finds the average quantity of the defined dimensions vertically and $WMAvg$ selects the weighted moving average quantity of the defined dimensions vertically.

Using the $V_{Avg}$ algorithm, we find the average quantity of the selected dimension vertically. We illustrate how the $V_{Avg}$ algorithm works in Table 1. Here we use only two dimensions: $Times$ dimension for a week’s sales only, and $Products$ dimension for only Beer and Bread. $Times$ dimension works as a grouping dimension. So the attributes of $Products$ dimension will be grouped according to its time frames (see Figure 2 for the detail of those dimensions).
Feature Selection for Web Page Classification

www.igi-global.com/chapter/feature-selection-web-page-classification/29152?camid=4v1a

Aspects of Multilingual News Summarisation

www.igi-global.com/chapter/aspects-of-multilingual-news-summarisation/96750?camid=4v1a

Raising, to Enhance Rule Mining in Web Marketing with the Use of an Ontology
Xuan Zhou and James Geller (2008). Data Mining with Ontologies: Implementations, Findings, and Frameworks (pp. 18-36).

www.igi-global.com/chapter/raising-enhance-rule-mining-web/7570?camid=4v1a